Hans Journal of Data Mining 数据挖掘, 2022, 12(2), 182-191
Published Online April 2022 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2022.122019

Hans汉斯

基于XGBoost等分类算法对葡萄酒数据集的R语言实践

王紫曦

东北大学秦皇岛分校数学与统计学院，河北 秦皇岛

收稿日期：2022年3月12日；录用日期：2022年4月12日；发布日期：2022年4月20日

摘 要

本文利用R语言研究了决策树、随机森林、支持向量机几种机器学习分类算法在葡萄酒数据集上的表现，分别得到的准确率为67.24%、68.15%、66.25%，表现较好。其中随机森林算法在这三种分类算法中表现最为良好，支持向量机表现最差，但三种算法的效果相差不大。由于三种算法在准确率上仍有较高的提升空间，因此引入极端梯度提升树(XGBoost)进行分类，该算法在随机森林与决策树的基础上进行改进，所得效果最好，为73.59%。然而直接基于R语言中四种机器学习算法对葡萄酒数据集分类所得效果较一般，仍需要在这个基础上予以改进。

关键词

逻辑回归，随机森林，决策树，支持向量机，极端梯度提升树

R Practice on Wine Datasets Based on Classification Algorithms Such as XGBoost

Zixi Wang

School of Mathematics and Statistics, Northeastern University at Qinhuangdao, Qinhuangdao Hebei

Received: Mar. 12th, 2022; accepted: Apr. 12th, 2022; published: Apr. 20th, 2022

Abstract

In this paper, we studied the performance of several machine learning classification algorithms of decision tree, random forest, and support vector machine on wine dataset using R language, and

文章引用: 王紫曦. 基于 XGBoost 等分类算法对葡萄酒数据集的 R 语言实践[J]. 数据挖掘, 2022, 12(2): 182-191.
DOI: 10.12677/hjdm.2022.122019

the accuracy obtained was 67.24%, 68.15%, and 66.25%, respectively, which performed well. The random forest algorithm performed the best among these three classification algorithms, and the support vector machine performed the worst, but the results of the three algorithms were not very different. Since there is still room for higher improvement in the accuracy of the three algorithms, the extreme gradient boosting tree (XGBoost) is introduced for classification, which improves on the random forest and decision tree and obtains the best result of 73.59%. However, the classification of the wine dataset directly based on the four machine learning algorithms in R is not very good and still needs to be improved on this basis.

Keywords

Logistic Regression, Random Forests, Decision Trees, Support Vector Machines, Extreme Gradient Boosting Trees

1. 引言

机器学习算法是一类能从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。其中分类算法是机器学习诸多算法中一个重要分支。与回归问题相比，分类问题的输出不再是连续值，而是离散值，用来指定其属于哪个类别。分类问题在现实中应用非常广泛，比如垃圾邮件识别，手写数字识别，人脸识别，语音识别等。

逻辑回归(Logistic Regression)、随机森林(Random Forest)、决策树(Decision Tree)、支持向量机(Support Vector Machine)、极端梯度提升树(Extreme Gradient Boosting)是分类问题中几种常用的算法。这几种算法已经被广泛运用于金融、电信、物流等多种行业内。同时，现有研究也在这几种分类算法上进行了优化，效果已经得到极大的提升。

统计显示，约 70%的开发人员使用 Python 语言来创建机器学习项目，仅 20%的开发人员使用 R 语言来创造机器学习项目。Python 专注于先验的准确性，因此在机器学习领域建立了良好的声誉。R 语言作为一种用于事实推理和静态推理的语言，在数据分析中得到了广泛的应用。R 语言是在统计分析与可视化中常用的优秀工具。随着机器学习与统计学之间的关系越来越密切，R 语言也出现越来越多的相关开源库，使得用 R 语言来创建机器学习项目越来越方便快捷。

本文中，作者利用 R 语言，基于决策树、随机森林、支持向量机、极端梯度提升树等算法在葡萄酒数据集上进行实验并比较了不同分类算法的效果。

2. 葡萄酒数据集和模型介绍

2.1. 模型介绍

随机森林(Random Forest)、决策树(Decision Tree)、支持向量机(Support Vector Machine)是分类问题中几种常用的算法。

1) 决策树

通常采用以下步骤进行决策树的构建[1]。

Step1：特征选择。首先选取一个典型特征作为划分标准。评估划分标准选取的不同，会导致选取特征的差异，由此形成出多种具有差异的决策树选择过程。

Step2：决策树的生成。将上述选取特征作为一定的评估准则，按照自上而下的顺序确定子节点。若数据集合内仅包含单一类别样本时，则此树终止向下分支。

Step3：剪枝决策树。在树的构建过程中容易出现过拟合问题，多数情况下的处理方法是进行剪枝。剪枝又包括预剪枝与后剪枝两类方法。

将特征空间表示为 X，假设输入向量 $X \in \mathbf{X}$， $X = (X_1, X_2, \cdots, X_p)$ 包含分类和有序类型的特征。将节点用 t 表示。有左子节点 t_L 和右子节点 t_R 。用 T 表示树中所有节点的集合，用 T 表示所有叶节点的集合 $\tilde{T}$ 。将拆分用 s 表示，分割集由 S 表示。

则决策树损失函数可表示为：

$$Ca(T) = \sum_{t=1}^{|T|} N_t H_t(T) + \alpha |T| \tag{1}$$

其中经验熵的计算公式为：

$$H(t) = -\sum_k \frac{N_{tk}}{N_t} * \log \frac{N_{tk}}{N_t} \tag{2}$$

将经验熵代入，可以得到：

$$Ca(T) = C(T) + \alpha |T| \tag{3}$$

2) 随机森林

随机森林(Random Forest)，是一种以选用决策树为基分类器的集成算法，组合多棵独立的决策树后，根据 Bagging 集成的方法得到最终结果。CART 算法一般是随机森林内部的决策树算法，每一棵 CART 树，通过 CART 算法的学习规则进行学习，学习的数据集对原始样本集采用了行抽样和列抽样，增强了抗拟合和抗噪音能力[2]。

假设样本集大小为 N，使用自助抽样 N 次有放回地随机抽取 N 个样本，某一个样本整个抽取过程中都不被抽中地概率为：

$$\left(1 - \frac{1}{N}\right)^{\frac{1}{N}} \tag{4}$$

在某一棵树的节点 m 处某特征的基尼系数为 GI。如果选择该特征为划分属性向下分裂，分裂后，左分支的基尼系数为 GIL，右分支的基尼系数为 GIR，则：

$$VI_m = GI - (GIR + GIL) \tag{5}$$

该特征分裂了 k 次，计算这棵树上这个特征的重要程度为：

$$\sum_{i=1}^{k} VI_i \tag{6}$$

如果这个特征在整个随机森林中共有 n 棵树被用到，计算在整个随机森林中该特征的重要程度为：

$$\sum_{j=1}^{n} \sum_{i=1}^{k} VI_{ij} \tag{7}$$

然后，进行归一化处理。将每一个特征的重要程度用一个度量的指标表示。在训练随机森林的每一棵决策树时，都会产生袋外数据，将这部分输入训练过的每棵树中预测，并将预测误差记录下来，每棵

树的误差为 errOOBI。通过将噪声干扰随即添加到该部分数据的所有样本的特定特征中，可以再次计算预测误差 errOOB2。随机森林中有 N 棵树时，特征属性重要性如下：

$$\sum \frac{\text{errOOB2} - \text{errOOB1}}{N} \tag{8}$$

3) 支持向量机

支持向量机的提出基于统计学习理论和结构风险最小化准则，其最终的目的是根据输入数据样本寻找到一个最优分类超平面[3]。

支持向量机最优分类面的求解问题可转化为求数据样本分类间隔最大化的二次函数的解，关键是求得分类间隔最大值的目标解。

基本分类判别面方程如下：

$$\omega^{\mathrm{T}} x + b = 0 \tag{9}$$

对线性可分的样本集进行归一化处理，分类间隔表达式如下：

$$\frac{2}{\|\omega\|} \tag{10}$$

通过加入有效约束条件，引入拉格朗日乘子后，解得最优分类判别函数，相应的支持向量机分类函数表达式如下：

$$f(x) = \operatorname{sgn}\left\{ \sum_{i=1}^{k} a_i^* y_i K(x_i \cdot x) + b^* \right\} \tag{11}$$

2.2. 红酒数据集介绍

葡萄酒是一种成分复杂的酒精饮料，不同产地、年份和品种的葡萄酒成分不同，这也是导致质量差异过大的重要属性。该数据集含有 12 个变量，包含了葡萄酒的不同理化特征。

固定酸度、挥发性酸度和柠檬酸的数据中都存在异常值。如果这些异常值被消除，变量的分布可以被认为是对称的。残糖呈正偏分布；即使在消除异常值之后，分布仍将保持偏态。游离二氧化硫，密度，有一些异常值，但这些与其他的非常不同。大多数异常值都在较大的一侧。酒精具有不规则形状的分布，但没有明显的异常值。

作变量间的相关系数图如图 1。

基于对相关系数图的概览，发现部分变量之间的相关系数上没有显著的不同。固定酸度与柠檬酸、固定酸度与密度、游离二氧化硫与总二氧化硫它们之间相关系数很大，也就是说这些变量数据之间的相关性很强。同时，固定酸度与 PH 值的相关系数接近−1，也就是说两种变量呈负相关详细。

对相关系数图下三角区域增加平滑拟合曲线与置信椭圆。平滑拟合曲线通过对数据进行拟合，直观表现出两个变量的相关关系。置信椭圆是对置信区间的描述，其长轴表示两个变量中第一个变量的置信区域，短轴表示两个变量中第二个变量的置信区域。而根据下图，可以看出挥发性酸度、硫酸盐和酒精等在不同的质量下有着更显著的差异。然而，不同质量之间的差异还不足以直接用一个变量来确定所有质量。

在图 1，图 2 的基础上，增加了对数据的散点图的绘制与概率分布图。同时，可以点击来进行交互查询某一变量与其他变量的相关关系，对图 3 的交互图进行观察，可以直观地看到在不同酒精、挥发性酸度和硫酸盐水平下的不同质量，从而可以对需要进行分析的变量进行快速筛选与概览。

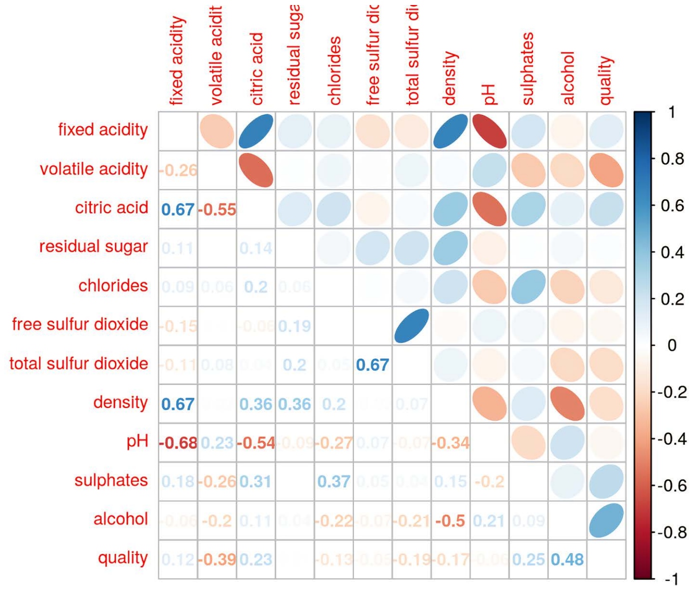

Figure 1. Correlation coefficient diagram of wine factors
图 1. 葡萄酒各属性相关系数图

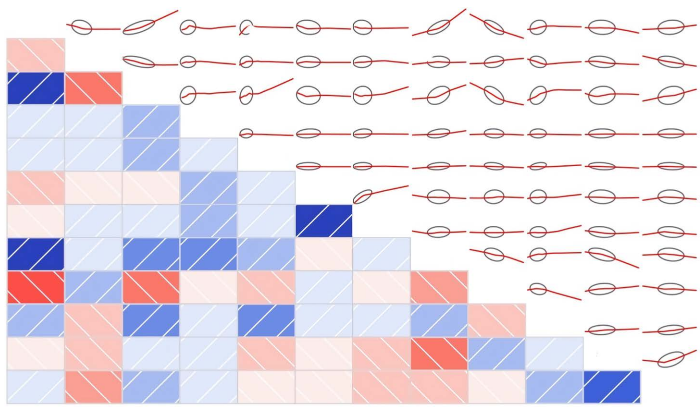

Figure 2. Correlation coefficient diagram of wine factors
图 2. 葡萄酒各属性相关系数图

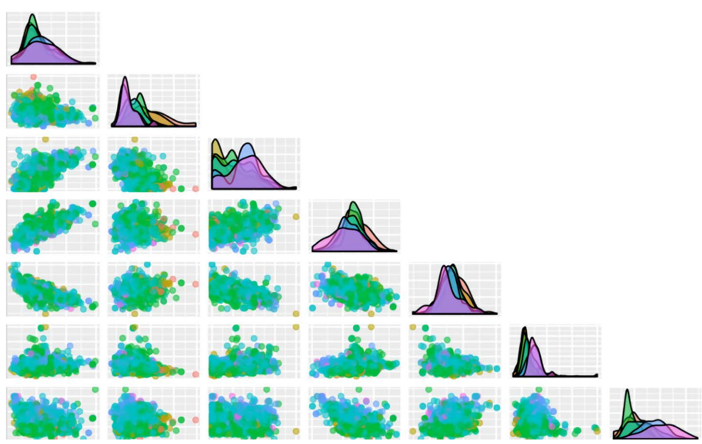

Figure 3. Interactive distribution trend chart of wine factors
图 3. 葡萄酒各属性交互式分布趋势图

2.3. 模型训练及其结果

导入葡萄酒数据集，对数据集进行预处理，删去数据集中异常值，然后对部分空缺值进行插值处理，删去部分空缺值。然后把预处理所得到的数据集按 7:3 的比例划分训练集与测试集。

直接调用决策树算法对葡萄酒数据集进行分类。决策树算法显示酒精、挥发酸和硫酸是确定质量的重要的变量。

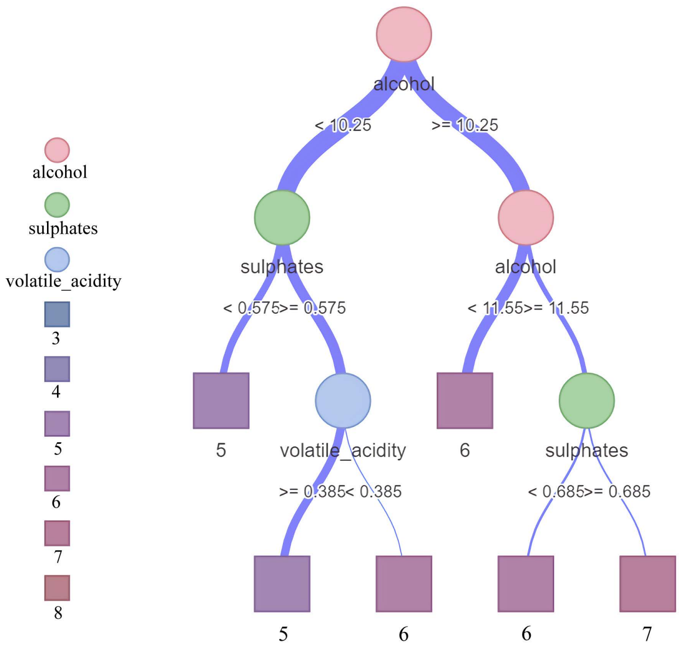

Figure 4. Decision tree classification results
图 4. 决策树分类结果

图 4 是基于红葡萄酒数据集建立的决策树，用于判断葡萄酒的质量好坏。葡萄酒主要具备三个属性：酒精含量、硫酸盐含量、挥发性酸度。每一个内部节点都表示一个属性判断，叶子节点表示葡萄酒质量等级。例如：葡萄酒含量低于 10.25，硫酸盐含量小于 0.575 的葡萄酒数据被划分成第五类。葡萄酒含量低于 10.25，硫酸盐含量高于 0.575 且挥发性酸度大于等于 0.405 的葡萄酒数据被划分成第五类。

Table 1. Significant factors obtained from the decision tree
表 1. 决策树所得重要属性

属性	总值
Alcohol	117.94
citric_acid	42.39
Density	45.50
pH	4.49
Sulphates	85.25
volatile_acidity	81.50

利用随机森林对葡萄酒数据集进行分类的结果如表 1 与决策树得到的结果如表 2 之间相差不大，均显示酒精、挥发性酸度和硫酸盐是确定质量的重要的变量。

利用随机森林算法对变量重要性排序如图 5，直观展现了变量数值大小关系，如 alcohol 数值最大，接近 200。

Table 2. Significant factors obtained from random forest
表 2. 随机森林所得重要属性

属性	总得分
Alcohol	192.15
citric_acid	159.27
Density	128.70
pH	159.94
Sulphates	131.67
volatile_acidity	154.44

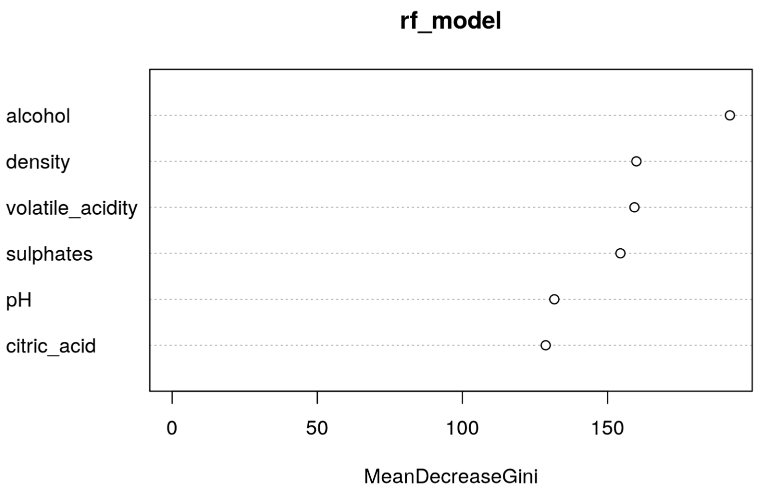

Figure 5. Scatterplot of variable importance under random forest algorithm
图 5. 随机森林算法下变量重要程度散点图

支持向量机也是分类时可以选择的方案之一，但是其在葡萄酒数据集上的效果并不突出，不如随机森林与决策树在葡萄酒数据集上的表现。

从表 3 中可以看出三种模型在该数据集上均有良好表现。但随机森林在该数据集上表现最好，准确率为 68.15%，三种数据集准确率相差不大。

Table 3. Accuracy performance of different algorithms on wine dataset
表 3. 不同算法在葡萄酒数据集上准确率表现

算法	准确率
决策树	67.24%
随机森林	68.15%
支持向量机	66.25%

3. 基于 XGBoost 的红酒数据集实验

由于随机森林、支持向量机、决策树在该数据集上表现仍存在很大改进空间[4]，在该部分作者采用极端梯度提升树对红酒数据集进行分类。

3.1. 模型介绍

极端梯度提升树是 Chen 等提出的基于集成思想的机器学习算法[5]。与传统的集成学习不同，传统的集成学习如随机森林是通过减少模型方差提高性能，而极端提升树是通过减少模型的偏差提高性能。作为机器学习算法的一种，XGBoost 在网络入侵检测、卫星网络协调态势预测等应用领域取得了良好的效果。极端提升树的主要思想就是基于当前的模型加入另一个模型，使得组合模型的效果优于当前模型[6]。

$$\widehat{y_j}=\sum_{j=1}^{n}f_l\left(x_j\right),f_l\in F \tag{12}$$

其中，$\widehat{y_j}$ 表示模型的预测值，L 表示树的数量，f_l 表示第 l 棵树模型，x_j 表示第 j 个输入样本，F 表示所有树模型的集合，则该模型的目标函数和正则项如公式(13) (14)所示：

$$Obj^{(t)}=\sum_{j=1}^{n}loss\left(y_j,\widehat{y_j}^{(t-1)}+f_t\left(x_j\right)\right)+\Omega\left(f_t\right)+c \tag{13}$$

$$\widehat{y_j}=\sum_{j=1}^{n}f_l\left(x_j\right),f_l\in F \tag{14}$$

其中 $Obj^{(t)}$ 表示构建第 t 棵树时的目标函数，$loss()$表示损失函数，一般为均方误差；$\widehat{y_i}^{(t-1)}$ 表示前 $t-1$ 棵树所计算的预测值；c 表示常数项；$\Omega\left(f_t\right)$ 表示第 t 棵树的正则项；γ 和 λ 表示正则项系数；T 表示某棵树所有叶子节点和数量；ω_o 表示某棵树中第 o 个叶子节点的权重。

$$Obj^{(t)}\simeq\sum_{j=1}^{n}\left[loss\left(y_j,\widehat{y_j}^{(t-1)}\right)+g_jf_t\left(x_j\right)+\frac{1}{2}h_jf_t^2\left(x_j\right)\right]+\Omega\left(f_t\right)+c \tag{15}$$

$$g_j=\partial_{\widehat{y_j}^{(t-1)}}loss\left(y_j,\widehat{y_j}^{(t-1)}\right) \tag{16}$$

$$h_j=\partial_{\widehat{y_j}^{(t-1)}}loss\left(y_j,\widehat{y_j}^{(t-1)}\right) \tag{17}$$

由于 $loss\left(y_j,\widehat{y_j}^{(t-1)}\right)$ 是固定值，因此可以并入常数项 c，而常数项对优化求解没有影响，因此可以去除，目标函数表示如公式(18)

$$Obj^{(t)}\simeq\sum_{j=1}^{n}\left(g_jf_t\left(x_j\right)+\frac{1}{2}h_jf_t^2\left(x_j\right)\right)+\Omega\left(f_t\right) \tag{18}$$

对目标函数进行变形

$$Obj^{(t)}=\sum_{o=1}^{T}\left[G_o\omega_o+\frac{1}{2}\left(H_o+\lambda\right)\omega_o^2\right]+\gamma T \tag{19}$$

求偏导后带入目标函数得到

$$Obj^{(t)}=-\frac{1}{2}\sum_{o=1}^{T}\frac{G_o^2}{H_o+T}+\gamma T \tag{20}$$

极端提升树利用贪心算法遍历树模型的所有分裂叶子节点，选择分裂后目标函数增益最大的叶子节点进行分裂，判定条件如下，大于 0 则可以分裂：

$$Gain=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{\left(G_L+G_R\right)^2}{H_L+H_R+\lambda}\right]-\gamma>0 \tag{21}$$

3.2. 模型训练结果

本实验对极端梯度提升树算法的参数作如下设定：

- booster:使用哪个弱学习器训练，默认 gbtree。
- eta(learning_rate)：learning_rate，在更新中使用步长收缩以防止过度拟合，设定为 0.08。
- gamma(min_split_loss)：设定为 0.7，分裂节点时，损失函数减小值只有大于等于 gamma 节点才分裂，gamma 值越大，算法越保守，越不容易过拟合，但性能就不一定能保证，需要平衡。
- alpha(reg_alpha)：默认= 0，权重的 L1 正则化项。增加此值将使模型更加保守。
- max_depth：设定为 8，一棵树的最大深度。增加此值将使模型更复杂，并且更可能过度拟合。
- min_child_weight：设定为 2，如果新分裂的节点的样本权重和小于 min_child_weight 则停止分裂。这个可以用来减少过拟合，但是也不能太高，会导致欠拟合。
- subsample：设定为 0.9，构建每棵树对样本的采样率。
- colsample_bytree：设定为 0.5，列采样率，也就是特征采样率。
- lambda(reg_lambda)：默认 = 1，L2 正则化权重项。增加此值将使模型更加保守。
- 任务参数 objective：设定为 multi:softmax：设置 XGBoost 以使用 softmax 目标进行多类分类，表示最小平方误差。
- eval_metric：设定为 merror，多类分类错误率验证数据的评估指标，将根据目标分配默认指标(回归均方根，分类误差，排名的平均精度)，用户可以添加多个评估指标。

利用极端梯度提升树计算所得分类时用以判断的属性特征重要性如表 4。

Table 4. Important factors of the red wine dataset obtained by XGBoost
表 4. XGBoost 所得红酒数据集的重要属性

属性	得分
Alcohol	0.15
Sulphates	0.13
volatile_acidity	0.11
total_sulfur	0.10
Density	0.10
Chlorides	0.08
pH	0.08

Table 5. Accuracy performance of different algorithms on wine dataset
表 5. 不同算法在葡萄酒数据集上准确率表现

算法	准确率
决策树	67.24%
随机森林	68.15%
支持向量机	66.25%
XGBoost	73.89%

从表 5 中可以看出与决策树、随机森林相比，该方法通过调整部分参数，准确率达到了 73.89%，提升效果显著。

4. 结论

决策树、随机森林、支持向量机与极端梯度提升树这几种算法，都有各自的优缺点，在实际应用中，我们应该根据不同的数据集的特点采用不同的模型。在红酒数据集中，极端梯度提升树算法的效果远远好于其他的算法。

而在不同算法适合的应用场景中，决策树与随机森林因其超强的学习能力更适合搜索排序的场景，支持向量机因其核函数敏感的特点适合高维文本分类、小样本分类，而 XGBoost 在多种应用场景中均表现良好。

同时，基于 R 语言的算法实现较为简洁，可视化的图更美观。因此，R 语言不仅可以用于传统的统计学习，更能在现有的机器学习与数据挖掘领域中发挥极大的作用。

参考文献

[1] 曲晨, 覃玉, 毛涛, 等. 决策树模型与 logistic 回归在中学生尝试吸烟影响属性中的应用[J]. 中国慢性病预防与控制, 2020, 28(4): 264-269.

[2] Liaw, A. and Wiener, M. (2002) Classification and Regression by Random Forest. *R News*, 23.

[3] Cauwenberghs, G. and Poggio, T. (2001) Incremental and Decremental Support Vector Machine Learning. *Advances in Neural Information Processing Systems*, **13**, 409-412.

[4] 杨剑锋, 乔佩蕊, 李永梅, 等. 机器学习分类问题及算法研究综述[J]. 统计与决策, 2019(6): 36-40.

[5] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203451

[6] 谭中明, 谢坤, 彭耀鹏. 基于梯度提升决策树模型的 P2P 网贷借款人信用风险评测研究[J]. 软科学, 2018, 32(12): 5.